EmoVIT: Revolutionizing Emotion Insights with Visual Instruction Tuning

Hongxia Xie[1], Chu-Jun Peng[2], Yu-Wen Tseng[2],
Hung-Jen Chen[3], Chan-Feng Hsu[3], Hong-Han Shuai[3], Wen-Huang Cheng[2]
[1]Jilin University
[2]National Taiwan University
[3]National Yang Ming Chiao Tung University

Abstract

Visual Instruction Tuning represents a novel learning paradigm involving the fine-tuning of pre-trained language models using task-specific instructions. This paradigm shows promising zero-shot results in various natural language processing tasks but is still unexplored in vision emotion understanding. In this work, we focus on enhancing the model's proficiency in understanding and adhering to instructions related to emotional contexts. Initially, we identify key visual clues critical to visual emotion recognition. Subsequently, we introduce a novel GPT-assisted pipeline for generating emotion visual instruction data, effectively addressing the scarcity of annotated instruction data in this domain. Expanding on the groundwork established by InstructBLIP, our proposed EmoVIT architecture incorporates emotion-specific instruction data, leveraging the powerful capabilities of Large Language Models to enhance performance. Through extensive experiments, our model showcases its proficiency in emotion classification, adeptness in affective reasoning, and competence in comprehending humor. The comparative analysis provides a robust benchmark for Emotion Visual Instruction Tuning in the era of LLMs, providing valuable insights and opening avenues for future exploration in this domain. Our code is available at https://github.com/aimmemotion/EmoVIT.

1. Introduction

Visual emotion recognition, a key area within artificial intelligence and computer vision, aims to predict human emotions based on visual cues such as facial expressions and body language. This technology is essential in bridging the gap between human affective states and machine understanding. Its diverse applications [10, 13, 22, 39], spanning from improving human-computer interaction to aiding in mental health assessment, underscore its significance. Accurate emotion recognition is vital for enhancing user experience and ensuring information security, as it helps prevent emotional manipulation and misinformation [32]. Developing robust emotion recognition models is not only a technical challenge but also a step towards more empathetic and intuitive AI systems, paving the way for more efficient and natural human-computer interactions.

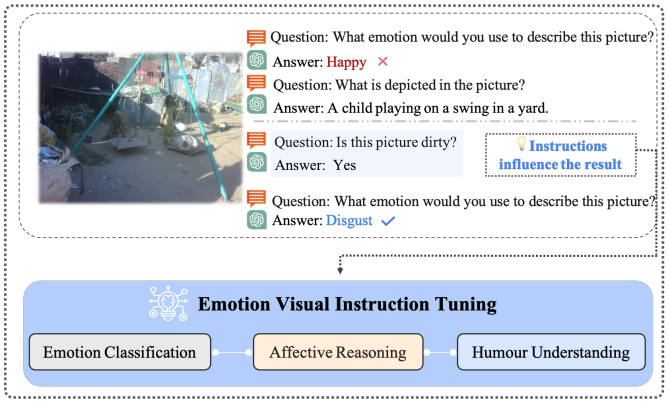


Figure 1. Illustration of the importance of instruction-following ability in visual emotion understanding.

The AI community has recently shown a growing interest in developing foundational vision models, *e.g.*, Flamingo [8], LLaVA [7], BLIP2 [14]. These models excel in open-world visual understanding, tackling several vision tasks such as classification, detection, segmentation, and captioning. In contrast, current large-scale multimodal models are still in its infancy when it comes to emotion perception [20]. As illustrated in Fig. 1, when directly query the GPT-4 [29] about the emotional category of an image, the model tends to provide incorrect responses. However, the model delivers accurate responses when provided with revised instructions. To fully leverage the potential of existing vision-based large models, our approach is based on the concept of *Instruction Tuning*. This effective strategy is aimed at teaching language models to follow natural language instructions, a technique proven to enhance their generalization performance across unseen tasks [7, 9, 21].

In this work, we focus on *developing the model's proficiency in understanding and following instructions related to emotional contexts*. This approach highlights the importance of fine-tuning the model's instruction-following capabilities, enabling it to interpret and respond to emotional content effectively. This is achieved by leveraging its pre-existing knowledge base, thereby eliminating the necessity for an emotion-specific architectural framework.

To address the notable challenges encountered in Instruction Tuning for visual emotion recognition, especially the lack of specific instruction data, we introduce a novel self-generation pipeline explicitly crafted for visual emotion recognition by using GPT-4 [29]. This innovative pipeline excels in generating a diverse array of (image, instruction, output) instances, thereby notably enhancing the dataset with a more extensive and task-oriented variety of examples. This approach not only overcomes the challenge of limited data availability but also reduces the dependence on human labor. Therefore, it streamlines the process, enabling more efficient and effective emotion recognition.

Additionally, Instruction Tuning has been criticized for its emphasis on surface-level features like output patterns and styles, rather than achieving a profound comprehension and assimilation of tasks [23]. To tackle this issue and enhance the diversity and creativity of instruction data, our dataset includes instructions that demand complex reasoning, going beyond basic question-and-answer formats. This is further enriched by incorporating visual cues such as *brightness*, *colorfulness*, *scene type*, *object class*, *facial expressions*, and *human actions*. These aspects are pivotal in fostering a nuanced comprehension of visual emotions, thus allowing the model to generate more precise and contextually appropriate interpretations [13].

After generating the emotion visual instruction data, we propose an Emotion Visual Instruction Tuning (EmoVIT) framework, leveraging the foundation of InstructBLIP [9]. This framework incorporates an emotion-centric, instruction-aware module that proficiently guides Large Language Models (LLMs) in assimilating the nuances of emotion instructions. Our work signifies a paradigm shift, presenting a new era of instruction-based learning for visual emotion understanding that relies less on explicit training data. Remarkably, as shown in Fig. 2, our approach requires almost 50% of the training data typically needed yet exceeds the performance of previous visual emotion recognition methods and popular Visual Instruction Tuning methods.

Our contributions can be summarized as follows:

- We explore the potential of the Visual Instruction Tuning paradigm for emotion comprehension and introduce the concept of Emotion Visual Instruction Tuning.
- After thoroughly considering the unique characteristics of visual emotion recognition, we develop a novel GPT-assisted pipeline for generating emotion visual instruction data. This approach effectively bridges the gap in available annotated instruction data within this specific domain.
- Building upon the foundation of InstructBLIP, our EmoVIT architecture integrates emotion domain-specific instruction data, harnessing the robust capabilities of LLMs to boost performance. The extensive experiments demonstrate our model's proficiency in emotion classification, affective reasoning, and comprehension of humour.

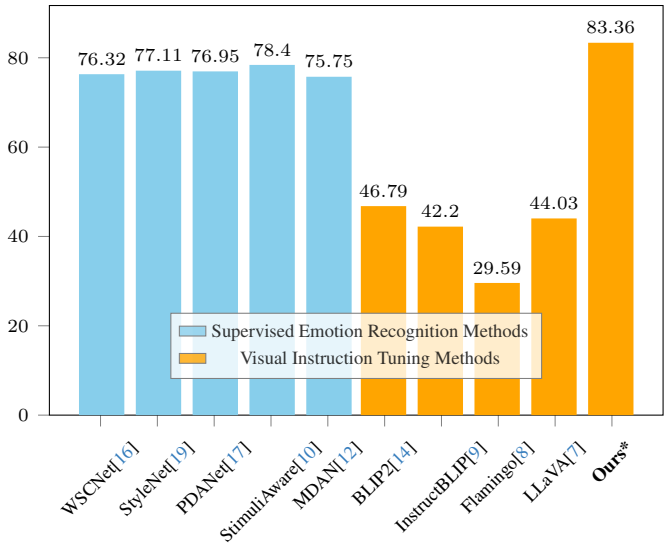


Figure 2. Performance comparison on EmoSet test set [13] (Accuracy %).

2. Related Work

2.1. Visual Emotion Recognition

A key challenge in visual emotion recognition is bridging the gap between an image's visual cues and the emotions it portrays [11, 12, 35]. While traditional efforts, *e.g.*, Xu *et al.*'s multi-level dependent attention network [12], focus on visual models for emotional feature learning, recent advancements like EmoSet [13] offer rich emotion-laden datasets with 3.3 million images. The rise of multimodal models, such as the GPT series [29], has further propelled Vision-Language Recognition. However, fully leveraging these models in emotion recognition is an area ripe for exploration. Our work leads the way in utilizing large-scale models for Emotion Visual Instruction Tuning.

2.2. Visual Instruction Tuning

Current Large Language Models (LLMs) have extensive knowledge bases, but their effectiveness depends on accurately interpreting human instructions due to a mismatch

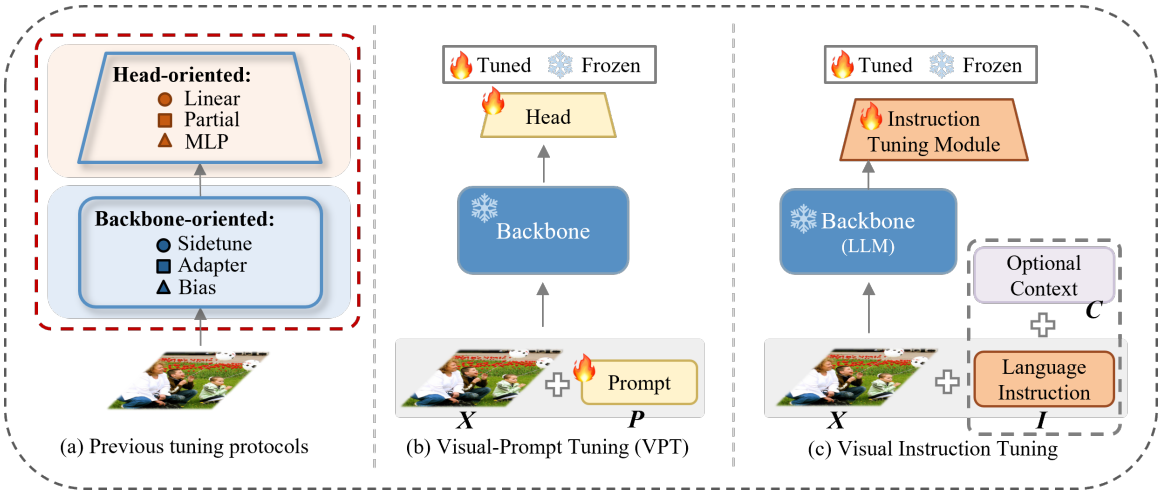


Figure 3. The comparison of different visual tuning paradigms.

between training goals and user expectations. LLMs are trained to minimize prediction errors, whereas users expect helpful and safe instruction-following. Instruction Tuning addresses this by teaching models to follow natural language instructions, enhancing generalization to new tasks. FLAN [21] demonstrated that training a large model on instruction-based datasets improves zero-shot performance. This approach has extended to vision-language tasks, with BLIP2 [14] and LLaVA [7] adapting instruction-tuned LLMs for visual inputs. InstructBLIP [9] introduces instruction-aware visual feature extraction and the Q-Former, enabling more flexible, instruction-driven feature extraction.

As a novel area, visual emotion instruction tuning lacks benchmarks or guidelines for creating emotion instruction data. Our work pioneers the use of large-scale models to develop an emotion instruction data pipeline, overcoming the limitations of manual annotation.

3. Method

3.1. Preliminary of Visual Instruction Tuning

In the deep learning era, **visual tuning** has experienced significant paradigm shifts, as depicted in Fig. 3.

In Fig. 3(a), conventional tuning methodologies encompass *Full fine-tuning, Head-oriented, and Backbone-oriented* techniques, capitalizing on large-scale pre-trained models. Predominantly, thoroughly fine-tuning these models for specific tasks, conducted end-to-end, is recognized as a highly effective strategy. However, this method requires maintaining separate copies of the backbone parameters for each distinct task, posing challenges in storage and deployment.

Alternatively, Visual Prompt Tuning (VPT) [24], presents an efficient substitute for full fine-tuning within large-scale vision Transformer models. It achieves this by employing a minimal fraction of trainable parameters in the input space while maintaining a frozen backbone model. The objective function for Visual Prompt Tuning is given by:

$$\min_{\theta_{\mathrm{P}}} \mathcal{L}(f(X, P; \theta_{\mathrm{P}}), Y) \tag{1}$$

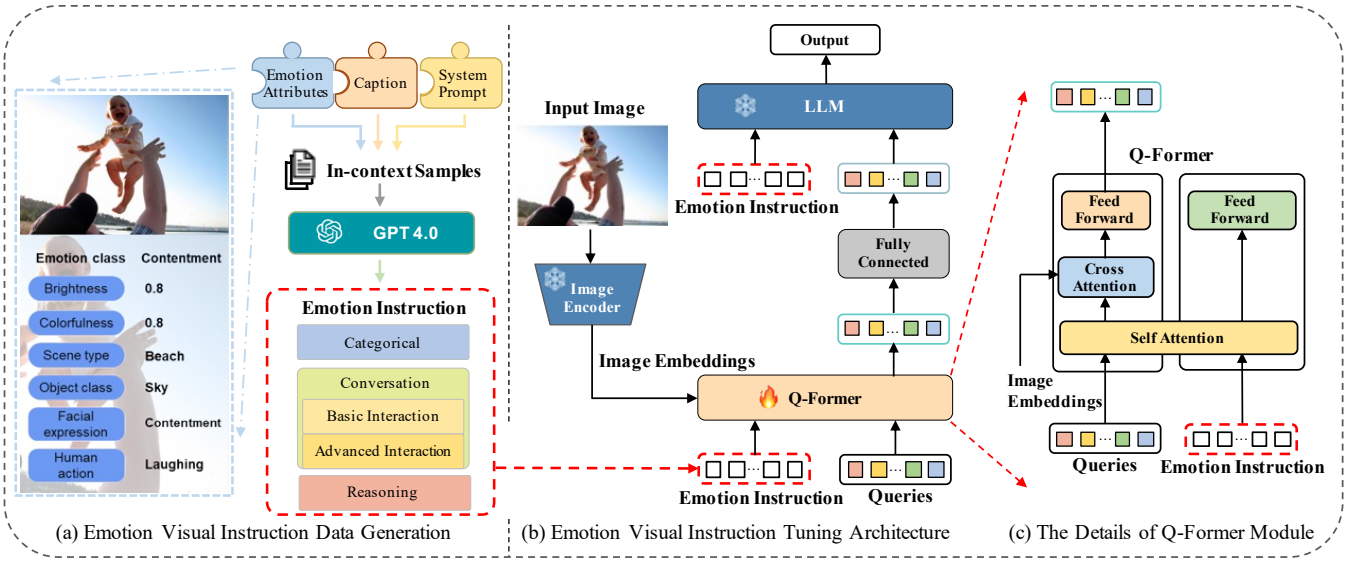
where $\min_{\theta_{\mathrm{P}}}$ is the minimization over the prompt parameters P, $\mathcal{L}$ is the loss function, f represents the model function with input image X, prompt parameters P, and learnable model parameters θ_{P} as input, and Y is the target output.

Visual Prompt Tuning focuses on optimizing LLMs using a small set of parameters, whereas Visual Instruction Tuning (VIT) aims to improve the model's comprehension of instructions, thereby addressing the model's shortcomings in specific domains. This type of method aims to enhance the model's proficiency in following instructions, leveraging the capabilities of the latest foundation models, *e.g.*, Llama [25], and BLIP2 [14]. Instructions serve as guiding constraints, shaping the model's outputs to conform to specific response characteristics and domain-relevant knowledge. This approach enables human monitoring of the model's behavior, thereby assuring alignment with the desired outcomes. Moreover, Instruction Tuning is computationally efficient, allowing LLMs to swiftly adapt to particular domains without extensive retraining or architectural alterations.

The objective function for Visual Instruction Tuning is given by:

$$\min_{\theta_{\text{tunable}}} \mathcal{L}(g(X, I, C; \theta_{\text{tunable}}), Y) \tag{2}$$

where $\min_{\theta_{\text{tunable}}}$ denotes the minimization over the tunable parameters θ_{tunable} in the Instruction Tuning Module, $\mathcal{L}$ is the loss function, g is the model function with instruction I, image X, other contexts C, and tunable parameters θ_{tunable},



(a) Emotion Visual Instruction Data Generation (b) Emotion Visual Instruction Tuning Architecture (c) The Details of Q-Former Module

Figure 4. The overall architecture of our proposed method. The *Emotion Instruction* data generated by (a) will be used for *Emotion Visual Instruction Tuning* in (b). During *Emotion Visual Instruction Tuning*, given an input image, the frozen Image Encoder initiates the process by extracting visual features. Emotion Instruction generated by (a) are subsequently interacting with *Queries* embedding through the learnable Q-Former. This interaction is key to drawing out image features that are relevant to the task at hand. As a result, the frozen LLM receives visual information conducive to instruction following.

and Y denotes the target output. The optional context C is not just raw data; it encompasses descriptive or directive information guiding the model on how to process input or which task to execute, *e.g.,* image caption. It's integral to the model's understanding and execution of tasks based on specific instructions or guidelines.

3.2. GPT-assisted Emotion Visual Instruction Data Generation

Previous methodologies commonly employed a consistent template-based set of instructions for every image within a dataset across various specific tasks [9]. For instance, a standard instruction such as "*Briefly describe the content of the image*" was employed uniformly across all images for Image Captioning. In this way, the model may not be able to adequately capture the unique characteristics of each image. Moreover, this *one-size-fits-all* approach often leads to suboptimal performance in emotion recognition tasks that require nuanced perception and differentiation of ambiguous emotion classes.

Since the topic of Emotion Visual Instruction Tuning is still in its infancy, no benchmarks or guidelines have been proposed so far for constructing emotion instruction data. Based on the recent successes of machine-generated instructions demonstrated in LLaVA [7], our work pioneers the use of existing LLMs to create a pipeline for self-generating emotion instructions. Different from previous template-based and *one-size-fits-all* instruction data, we propose an instance-wise and LLM-assisted visual emotion instruction data pipeline. This methodology transcends the constraints of manual annotation by employing GPT-4 [29] to generate instance-wise, tailored instruction data that dynamically corresponds to visual content.

Prior to the development of instructional data for the visual emotion recognition task, it is imperative to confront a fundamental academic problem: *What types of visual clues are pivotal in identifying emotions?* This necessitates a careful consideration of the unique characteristics inherent to the task, along with a comprehensive understanding of the potential visual cues associated with human emotions. In this work, we propose a novel visual instruction data mechanism to remove the inherent subjectivity and ambiguity in emotional interpretation. Specifically, we integrate a broad spectrum of emotion attributes across multiple levels: *low-level* attributes (*e.g.*, brightness, colorfulness), *mid-level* attributes (*e.g.*, scene type and object class), and *high-level* attributes (*e.g.*, facial expressions and human actions), building upon insights from previous work [13]. This comprehensive strategy not only aligns with the intricate nature of emotions but also significantly enhances the model's capability to interpret and understand visual emotional cues more accurately and holistically.

The overall pipeline of our proposed emotion visual instruction data is shown in Fig. 4 (a). For an image X_{img}, three types of image-related contexts are essential for GPT-4 to generate emotion instruction data: (i) a caption X_c, (ii) an emotion attribute list X_{attr}, which includes *emotion class, brightness, colorfulness, scene type, object class, facial expression*, and *human action*, and (iii) the system prompt, designed to enable GPT-4 to comprehend the specific task

requirement[1].

We first manually design a few examples which are used as seed examples for in-context learning to query GPT-4. This operation leverages the model's ability to extrapolate from given examples, enhancing its understanding and response accuracy based on the principles of few-shot learning [7]. Our generated emotion instruction data includes three types: *Categorical*, *Conversation*, and *Reasoning*. Building upon previous research [7], our generated instruction data adheres to the dialogue format, exemplified in Fig. 5.

Our strategy for generating emotion instruction data adopts a progressive approach *from simple to complex*. Initially, for the *Categorical* data, we transform the associated emotion class of the image into a structured format. This process serves as the foundational component of our emotion instruction data.

For the *Conversation* data, our framework is designed to create dialogues in which the GPT assistant interacts with an inquirer, focusing on the emotion attributes of the image. In this setup, the assistant's responses are tailored to interpret and describe the image as though it were within its own visual field, thereby providing insights from an observational viewpoint. The scope of questions posed is comprehensive, encompassing the types of objects depicted, their actions, and the dynamics of their interrelationships. The dialogues we generate fall into two categories: (i) *Basic Interaction*, focusing on the provided emotion attribute list with simple, direct characteristics, and (ii) *Advanced Interaction*, which builds on the first type to reach greater conversational complexity and sophistication.

For the *Reasoning* data, our approach extends beyond mere visual content, prompting the model to generate in-depth reasoning questions. To enhance the dialogue's credibility and structure, detailed examples are incorporated alongside logical reasoning steps, ensuring that the discourse convincingly captures the intricacies of the visual content.

3.3. Emotion Visual Instruction Tuning

After acquiring the emotion visual instruction data as detailed in Sec. 3.2, our goal is to employ this data in enhancing the existing Visual Instruction Tuning model. This enhancement aims to align the LLMs' existing knowledge with the emotion understanding domain.

As shown in Fig. 4 b, we have developed an Emotion Visual Instruction Tuning (EmoVIT) architecture based on InstructBLIP [9]. This architecture specifically leverages its Instruction-aware Q-Former Module, as depicted in Fig. 4 c, for emotion-centric instructional tasks.

[1]A detailed description of the system prompt is provided in the supplementary materials.

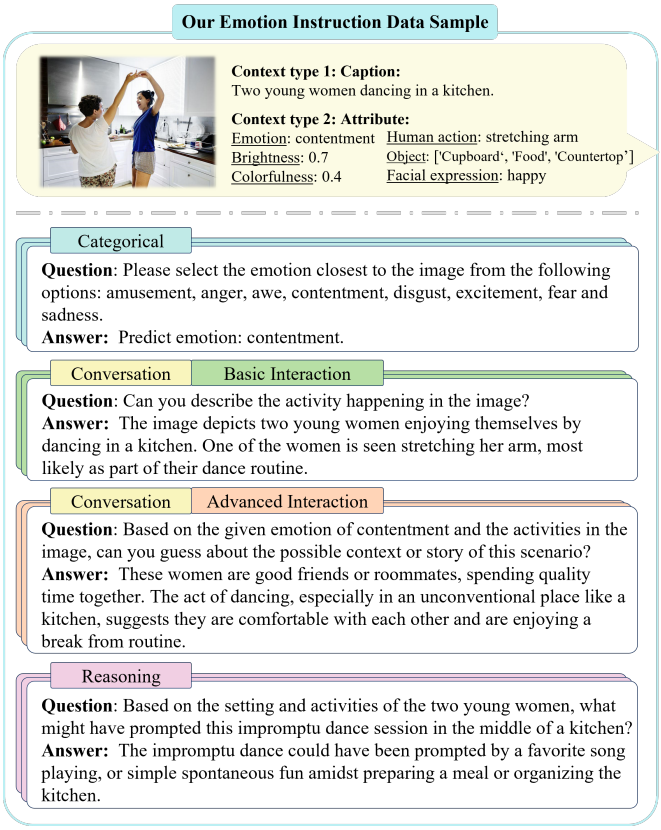


Figure 5. The sample of our generated visual emotion instruction data.

Specifically, the Instruction-aware Q-Former Module takes in the *emotion instruction tokens*, *queries*, and *image embeddings* as input. The image embeddings are extracted by a frozen image encoder. The learnable queries are initially produced by the pre-trained Q-Former of InstructBLIP. During training, the Instruction-aware module enhances task-specific feature extraction. It does this by integrating emotion instruction and query embeddings within self-attention layers, aligning visual information with the LLM's instruction-following requirements. Our approach adopts cross-entropy loss, tailoring it to the intricacies of visual emotion recognition tasks, thus ensuring precise and contextually relevant model training outcomes.

We note that the data generated by our approach is not confined to a single model but can also be applied to other Visual Instruction Tuning models, such as LLaVA [25]. Notably, when LLaVA is fine-tuned with our data, it exhibits a significant enhancement in emotion recognition capabilities, as detailed in Sec. 4.2. In this way, we demonstrate not only the effectiveness but also the transferability of our generated data, showing its broad applicability and impact.

4. Experimental Results

4.1. Implemental Details

Our implementation is based on the LAVIS library [31]. Our EmoVIT starts with a pre-trained InstructBLIP baseline and proceeds to fine-tune exclusively the Q-Former module, whilst keeping both the image encoder and the language model frozen. The parameters for our training adhere to the default settings established by InstructBLIP.

Datasets. We evaluate our framework on ten benchmark datasets annotated under different scenarios and class number, namely EmoSet [13], WEBEmo [11], Emotion6 [34], the Flickr and Instagram (FI) [35], Artphoto [36], IAPS [37], Abstract [36], EmotionROI [38], UnbiasedEmo [11], and OxfordTVG-HIC [33].

Held-in Pretraining. Following previous work [9], we divide our dataset into two categories: held-in for pretraining and held-out for evaluation [2]. Considering the EmoSet dataset's comprehensive inclusion of emotion attributes for each image, it has been chosen as the primary resource for our held-in pretraining phase. Simultaneously, for a broader assessment, we perform held-out evaluations using the test sets from various other datasets.

For the generation of emotion visual instruction data, we initially employ the BLIP2 model for image captioning, followed by leveraging the GPT-4 API to generate emotion instruction data. In total, our collection comprises *Categorical*, *Conversation*, and *Reasoning* instruction data, derived from 51,200 unique images. This represents less than 50% of the entire EmoSet.

4.2. Held-out Evaluation

As shown in Tab. 1, our proposed methodology exhibits a marked superiority in performance relative to the burgeoning Visual Instruction Tuning Methods. Since they have been pre-trained on dozens of large-scale datasets, it is evident that our generated emotion visual instruction data is particularly effective for emotional understanding Our results signify a paradigm shift, heralding a new era of model training that relies less on explicit supervision and more on the robustness of emotion instruction-driven learning.

The Effectiveness of Our Proposed Emotion Visual Instruction Data. As the first to introduce the concept of emotion visual instruction data, our study seeks to evaluate the generalizability of this newly generated instruction data. Our goal is to test its efficacy not only with InstructBLIP but also across other Visual Instruction Tuning model, to understand its broader applicability. As depicted in Fig. 6, we employ two Visual Instruction Tuning models, LLaVA and InstructBLIP, which were fine-tuned on our specially generated emotion visual instruction data. Subsequent testing across five distinct datasets reveals notable improvements in both models, substantiating the efficacy of our generated data. Notably, InstructBLIP demonstrated a more substantial overall enhancement compared to LLaVA. This can be attributed to InstructBLIP's specialized Instruction-aware Q-Former Module, which adeptly extracts the salient features of our emotion instructions and synergizes them effectively with the corresponding images, thereby yielding improved performance.

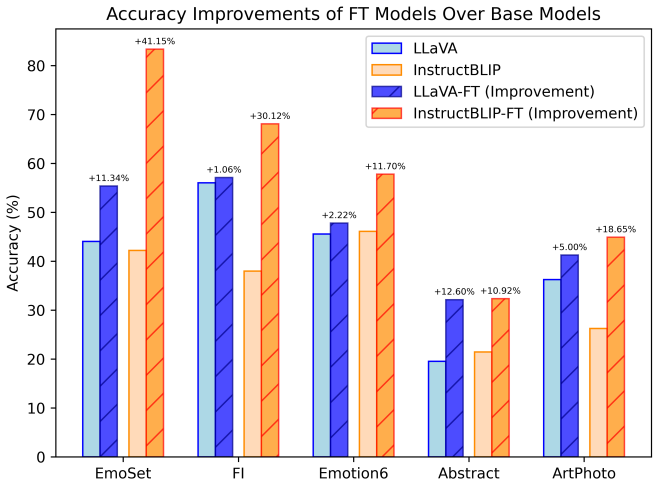


Figure 6. The improvement of our proposed emotion visual instruction tuning data tuning on LLaVA [7] and InstructBLIP [9].

4.3. Effectiveness of Different Instruction Data

4.3.1 Ablation Study of Different Instruction Data

The ablation study outlined in Tab. 2 provides a comprehensive analysis of the impact that different instructional data types have on model performance, specifically concerning accuracy metrics on the EmoSet test set. Initially, the model, referred to as InstructBLIP [9], operates without the integration of the three types of instructional data and attains a baseline accuracy of 42.20%. This foundational performance is significantly enhanced with the inclusion of *Categorical* data, which alone contributes to a substantial increase in accuracy. The introduction of *Conversation* data further amplifies this effect, underscoring the value of conversational context in improving the model's predictive capabilities. The addition of *Reasoning* data notably boosts performance, achieving a peak accuracy of 83.36%. This indicates that the model significantly benefits from the nuanced cues in reasoning, aiding in understanding complex emotional instructions. The gradual improvements with each data type support the idea that a diverse approach to instructional data markedly enhances model comprehension and performance.

[2]Unlike the setup in InstructBLIP, our dataset exclusively comprises emotion-related content. Consequently, our held-out evaluation does not constitute a strict zero-shot evaluation in the conventional sense.

Method	WebEmo	FI	Emotion6	Abstract	ArtPhoto	IAPSa	EmotionROI	EmoSet
Number of Classes	25	8	6	8	8	8	6	8
Flanmingo [8]	9.36	14.91	21.67	3.57	17.5	10.13	21.72	29.59
LLaVA [7]	12.55	56.04	49.44	19.54	36.25	42.43	46.46	44.03
BLIP2 [14]	20.10	57.72	50.00	28.57	36.25	39.24	50.51	46.79
InstructBLIP [9]	12.80	37.97	46.11	21.42	26.25	34.18	46.13	42.20
Ours*	**21.12**	**68.09**	**57.81**	**32.34**	**44.90**	**44.13**	**53.87**	**83.36**

Table 1. Held-out performance comparison on visual emotion datasets (%).

Categorical	Conversation	Reasoning	Accuracy (%)
-	-	-	42.20
✓	-	-	80.90 (+38.70)
✓	✓	-	81.95 (+39.75)
✓	✓	✓	83.36 (+41.16)

Table 2. Ablation study of three types of instruction data. Accuracy (%) on EmoSet test set.

4.3.2 Instruction Sensitivity

This work is dedicated to the creation of a varied corpus of visual emotion instruction data, alongside the development of a robust instruction-based model. Our objective is for the model to demonstrate stability, producing consistent results in the face of minor variations in instruction phrasing, provided the core objective of the task persists unchanged. To this end, we employ the *Sensitivity* evaluation metric, as introduced by [30], to assess the model's fidelity in generating uniform outcomes irrespective of instructional nuances.

We employ two semantically similar instructions as input prompts for the model, testing their impact on the Sensitivity score across three visual emotion datasets for different Visual Instruction Tuning models. The first instruction is: "From the given options: `cls_1`, `cls_2`, `cls_3`, etc., identify the emotion that most accurately reflects the image. Ensure your selection is confined to the listed options. Respond in the format: Predicted emotion:" The second one states: "Please choose the emotion that best corresponds to the image from the following options: `cls_1`, `cls_2`, `cls_3`, etc. (Do not provide answers beyond the provided candidates.) Please reply in the following format: Predict emotion:"

As illustrated in Fig. 7, our approach, along with BLIP2, exhibited exceptionally low Sensitivity values, demonstrating robustness in understanding the instructions. Conversely, Flamingo and InstructBLIP displayed a higher degree of sensitivity, indicating a relative susceptibility to variations in instruction wording.

4.4. Robustness

Given that current emotion recognition datasets often exhibit category imbalances and labeling biases, our aim is to evaluate the generalization ability of various learning strategies more impartially. Hence, we selected the UnBiasedEmo test set [11], which is uniquely suited for recognizing intricate emotions, such as those associated with identical objects or scenes, *e.g.*, landscapes, crowds, families, babies, and animals, where the emotional undertones can be particularly subtle and complex.

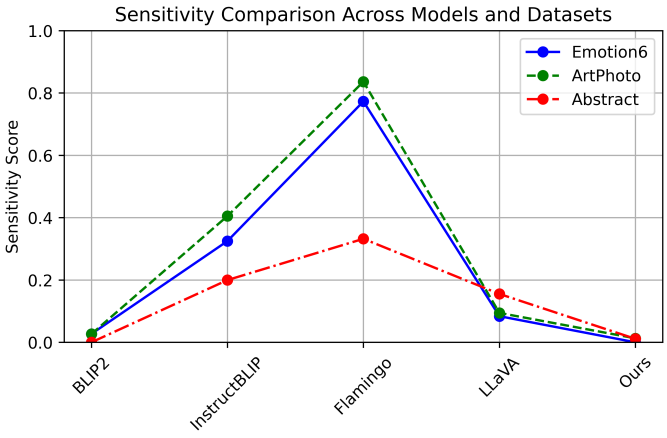


Figure 7. The sensitivity score comparison (the lower the better).

As depicted in Tab. 3, our proposed methodology demonstrates superior performance when benchmarked against conventional supervised emotion recognition techniques, thereby underscoring the efficacy of our approach in more accurately discerning complex emotional contexts.

Method	Accuracy (%)
Direct Learning [11]	71.64
Self-Directed Learning [11]	72.45
Joint Learning [11]	71.64
Curriculum Learning [11]	74.27
Ours*	**74.72**

Table 3. Performance comparison on UnbiasedEmo dataset.

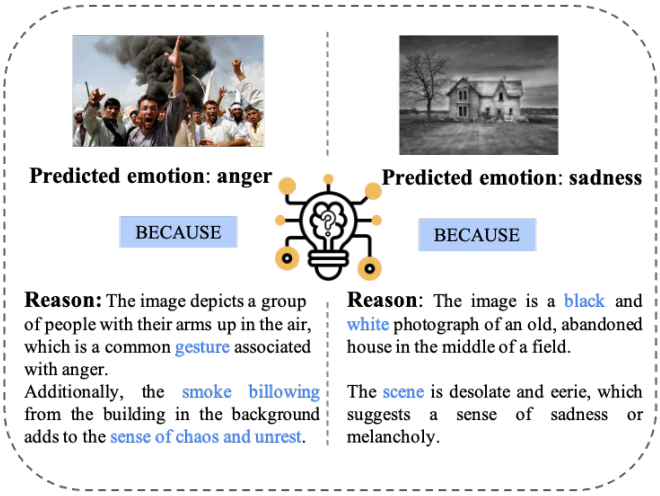


Figure 8. The sample of our generated explanation.

4.4.1 Affective Reasoning

In the domain of visual emotion recognition, where ambiguity and subjectivity are pervasive, the advent of an interpretable model is of considerable value. Such a model elucidates its cognitive processes, enhancing its trustworthiness and practicality in scenarios requiring a delicate grasp of emotional subtleties.

Leveraging Visual Instruction Tuning, our model transcends mere categorization of emotions; it articulates the underlying rationale for its classifications. The executing commands for identifying emotions and elucidating the decision basis is illustrated below:

```
Predicted emotion: [emotion].
Reason: [explanation].
```

Our model delineates the visual features influencing its determinations, thereby addressing the complexities inherent in discerning and explaining emotion-related nuances.

The explanations provide us with visual clues contained within the images, as exemplified in Fig. 8. It provides interpretable visual indicators that inform the model's outputs, as demonstrated in our example, by disambiguating the often abstract emotional categories.

4.5. Scaling Law

Pretraining data. As demonstrated in Tab. 4, there is a clear correlation between the size of the pre-training dataset and improved performance. Consequently, we anticipate that an increase in training data in the future could enhance the effectiveness of Emotion Visual Instruction Tuning.

4.6. Humour Caption Generation

The comprehension of humor is intricately linked to the understanding of emotions. Leveraging our generative language model, we conduct a caption generation task without modifying the model's architecture, specifically testing the model's proficiency in generating humorous captions. For this purpose, we select 50 images from the OxfordTVG-HIC dataset [33] and generate corresponding captions using our model. Subsequently, the captions produced by our model are compared with manually annotated captions from the dataset in a user study. Thirty participants were asked to vote on which captions were more humorous. Our model-generated captions receive 60% of the votes, demonstrating its effective humor generation capabilities. One sample is visualized in Fig. 9.

5%	10%	30%	50%
79.00	81.00	79.34	83.36

Table 4. Ablation study of different portion of pre-training data. Accuracy (%) on EmoSet test set.

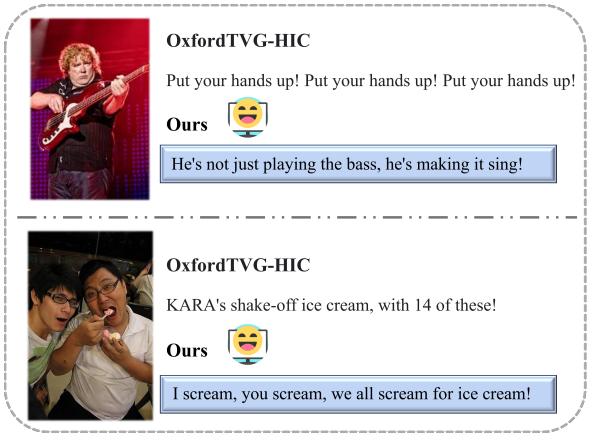


Figure 9. The sample of our generated humour caption *vs* human writing humour caption from OxfordTVG-HIC.

5. Conclusion

In our study, drawing upon the distinctive visual cues key to visual emotion recognition, we present a GPT-assisted pipeline specifically designed for generating emotion visual instruction data. The developed EmoVIT model incorporates emotion-specific instructions, leveraging LLMs for enhanced performance. Our comprehensive experiments validate its effectiveness in emotion classification, affective reasoning, and humor understanding. This comparative analysis sets a benchmark for Emotion Visual Instruction Tuning with LLMs, providing valuable insights and directions for future research in this field.

EmoVIT: Revolutionizing Emotion Insights with Visual Instruction Tuning

Supplementary Material

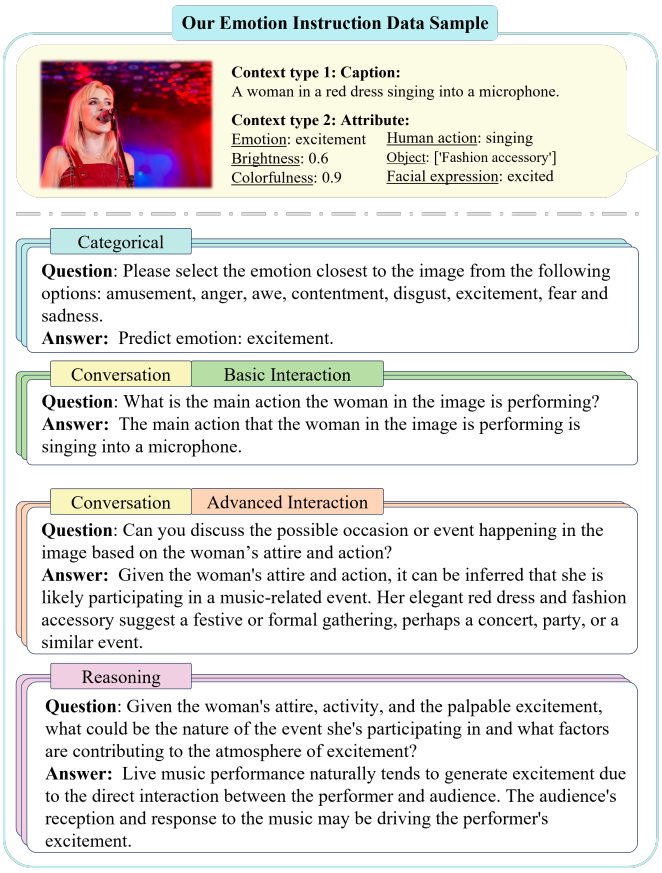


Figure 10. The sample of our generated visual emotion instruction data.

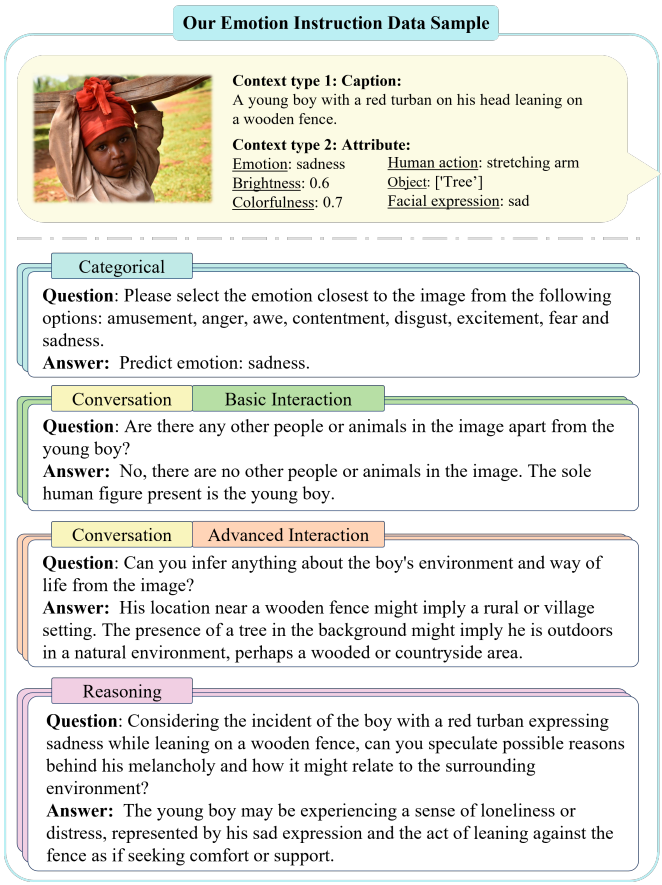


Figure 11. The sample of our generated visual emotion instruction data.

6. More Emotion Visual Instruction Data Sample

Additional samples from our Emotion Visual Instruction Data collection are presented in Figures 10 and 11. Upon acceptance, the complete dataset will be made available on our project webpage.

7. Implemental Details

7.1. Our Experiment Settings

Held-out vs *supervised learning*. We adopt the terminology *held-in* and *held-out* as defined in the work of InstructBLIP [9]. For the held-in, we utilize the training subset of the EmoSet dataset for Emotion Visual Instruction Tuning, with its corresponding test subset serving the purpose of held-in evaluation. The outcomes of this evaluation are depicted in Fig. 1 of the main manuscript.

In our held-out evaluation, we focus on determining how instruction tuning bolsters the model's ability to transfer learning to new and unseen data. It's crucial to highlight that our methodology sets a distinct path from InstructBLIP's framework. Our dataset is specifically curated with emotion-centric content, presenting unique categories such as *cheerfulness* and *enthrallment* found in WEBEmo, which are not typically included in other datasets. Conversely, common emotional categories like *anger* and *fear* are shared with other collections, such as FI and Emotion6. This distinctive mix in our dataset implies that our held-out evaluation operates on a *cross-domain* level, examining the model's ability to interpret and adapt to diverse emotional contexts not strictly confined to zero-shot scenarios.

7.2. System Prompt

The system prompt inputted into ChatGPT for the purpose of gathering instruction-based data is presented below.

You are an AI visual assistant, and you are seeing a single image. What you see are provided with one caption and some emotion related attributes, describing the same image you are looking at. Answer all questions as you are seeing the image. The range of brightness is from 0 (darkest) to 1 (brightest), and the range of colorfulness is from 0 (black-and-white) to 1 (the most colorful).

Design two questions for a conversation between you and a person asking about this photo. The answers should be in a tone that a visual AI assistant is seeing the image and answering the question. Ask diverse questions and give corresponding answers.

Include questions asking about the visual content of the image, including the object types, object actions, relationship among objects, etc. Only include questions that have definite answers: (1) one can see the content in the image that the question asks about and can answer confidently; (2) one can determine confidently from the image that it is not in the image. Do not ask any question that cannot be answered confidently. Please answer with the format Question: Answer:

Also include one complex question that is relevant to the content in the image, for example, asking about background knowledge of the objects in the image, asking to discuss about events happening in the image, etc. Again, do not ask about uncertain details. Provide detailed answers when answering complex questions. For example, give detailed examples or reasoning steps to make the content more convincing and well-organized. You can include multiple paragraphs if necessary.

7.3. Details of the Q-Former

Similar to the approach in InstructBLIP, Q-Former is a lightweight transformer architecture that utilizes a collection of trainable query vectors to distill visual features from a static image encoder. The Q-Former acts as the trainable module to bridge the gap between a frozen image encoder and a frozen LLM. Its role is to curate and present the most pertinent visual information, thereby enabling the LLM to generate the targeted textual output efficiently. Following the default setting, in our experimental setup, we employ 32 distinct queries, each with a dimensionality of 768.

7.4. Sensitivity Formula

As mentioned in Sec.4.3.2 in the main paper, we employ the *Sensitivity* evaluation metric, as introduced by [30], to assess the model's fidelity in generating uniform outcomes irrespective of instructional nuances. Specifically, for each task $t \in T$, given its associated instances with task instructions: $D^t = \{(I_j^t, x_j^t, y_j^t) \in T \times X^t \times Y^t\}_{j=1}^N$, sensitivity is defined as:

$$\mathbf{E}_{t \in T} \left[\frac{\sigma_{i \in I^t} \left[\mathbb{E}_{(x,y) \in D^t} \left[\mathcal{L}(f_\theta(i,x),y) \right] \right]}{\mu_{i \in I^t} \left[\mathbb{E}_{(x,y) \in D^t} \left[\mathcal{L}(f_\theta(i,x),y) \right] \right]} \right] \tag{3}$$

where $\mathcal{L}$ denotes the evaluation metric, *i.e.*, emotion classification accuracy, $f_\theta(\cdot)$ represents the Visual Instruction Tunign model. The standard deviation and mean of the model's performance across all instructions are denoted by $\sigma_{i \in I^t}[\cdot]$ and $\mu_{i \in I^t}[\cdot]$, respectively.

8. Ablation Study of LLM Model Size

In our attempts with the EmoVIT architecture's LLM, we explored the use of models of varying sizes (as shown in Tab. 5). The results indicated that the smaller model, Vicuna7B, outperformed its larger counterparts. This may be attributed to the limited training data available for our task, which potentially underutilizes the capabilities of larger models. Consequently, we anticipate that an increase in training data in the future could enhance the effectiveness of Emotion Visual Instruction Tuning.

Vicuna-7B	Vicuna-13B	FlanT5XL
83.36	82.21	80.98

Table 5. Ablation study of different LLM model size. Accuracy (%) on EmoSet test set.

9. GPT-4 *vs* GPT-4 Turbo

We conducted a comparative analysis of conversational datasets derived from GPT-4 (the model name is *gpt-4* in the API) against the recently released GPT-4 Turbo (the model name is *gpt-4-1106-preview* in the API). The comparative metrics yielded negligible differences between the two models (83.36% *vs* 82.96% on EmoSet test set).

10. Adding In-context Samples in Held-out Evaluation

Recent LLMs are capable of in-context learning when provided with a limited number of examples in a few-shot manner. In this work, we have also embarked on such an exploration. For instance, Tab. 6 presents the in-context samples utilized within the EmotionROI dataset. During our held-out evaluation, we incorporated three in-context samples for each category, consisting of a caption paired with its corresponding emotion class. Nevertheless, in our experimental observations, we did not witness any enhancement in performance attributable to furnishing the LLM with these in-context examples. Consequently, our finalized methodology did not incorporate in-context samples during the held-out evaluation phase.

Description	Emotion
Unleashed Fury: A portrait of raw, unfiltered anger etched on the subject's face.	Anger
Volcanic Eruption in Human Form: A Portrait of Unrestrained Fury.	Anger
An explosive portrait of raw fury, where every clenched jaw and furrowed brow tells a tale of unchecked anger.	Anger
Face contorted in a grimace of pure disgust, as if they just tasted a year-old lemon.	Disgust
Caught in the throes of revulsion, a face grimaces as if it just tasted the world's sourest lemon.	Disgust
Picture Perfect: A Masterclass in the Art of Disgust Expression	Disgust
A chilling moment of pure terror, etched in every detail.	Fear
A chilling moment of pure terror etched on the face, a stark embodiment of fear.	Fear
someone with a wide smile, a group	Joy
Overflowing with joy, like a puppy at a park!	Joy
A poignant portrait of sorrow, where teardrops are the silent language of grief.	Sadness
An evocative portrayal of sorrow, with shadows seemingly swallowing the light, reflecting the heavy weight of sadness.	Sadness
An abstract portrayal of solitude, where the vivid hues of melancholy paint a poignant picture of sadness.	Sadness
Caught in a moment of pure astonishment, eyes wide and mouth agape.	Surprise
Caught in the headlights of astonishment: a jaw-dropping moment of surprise!	Surprise
Caught in the Act! A person's wide-eyed gasp of sheer surprise.	Surprise

Table 6. Illustrative Examples of Emotion Descriptors in Visual Data

11. Limitation and future work

Due to the reliance on the GPT-API and cost considerations, our held-in pretraining phase utilized less than 50% of the EmoSet dataset. Despite outperforming other methods, we recognize the potential for significant improvements in future work by expanding the data scale. We anticipate that advancements in visual emotion understanding will parallel increases in both data and model scale.

References

[1] F. LastName, "The frobnicatable foo filter," 2014, face and Gesture submission ID 324. Supplied as supplemental material `fg324.pdf`.

[2] ——, "Frobnication tutorial," 2014, supplied as supplemental material `tr.pdf`.

[3] F. Alpher, "Frobnication," *IEEE TPAMI*, vol. 12, no. 1, pp. 234–778, 2002.

[4] F. Alpher and F. Fotheringham-Smythe, "Frobnication revisited," *Journal of Foo*, vol. 13, no. 1, pp. 234–778, 2003.

[5] F. Alpher, F. Fotheringham-Smythe, and F. Gamow, "Can a machine frobnicate?" *Journal of Foo*, vol. 14, no. 1, pp. 234–778, 2004.

[6] F. Alpher and F. Gamow, "Can a computer frobnicate?" in *CVPR*, 2005, pp. 234–778.

[7] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," in *Advances in Neural Information Processing Systems*, 2023. 1, 2, 3, 4, 5, 6, 7

[8] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 23 716–23 736. 1, 2, 7

[9] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, "Instructblip: Towards general-purpose vision-language models with instruction tuning," in *Advances in Neural Information Processing Systems*, 2023. 1, 2, 3, 4, 5, 6, 7

[10] J. Yang, J. Li, X. Wang, Y. Ding, and X. Gao, "Stimuli-aware visual emotion analysis," *IEEE Transactions on Image Processing*, vol. 30, pp. 7432–7445, 2021. 1, 2

[11] R. Panda, J. Zhang, H. Li, J.-Y. Lee, X. Lu, and A. K. Roy-Chowdhury, "Contemplating visual emotions: Understanding and overcoming dataset bias," in *Proceedings of the European Conference on Computer Vision*, 2018, pp. 579–595. 2, 6, 7

[12] L. Xu, Z. Wang, B. Wu, and S. Lui, "Mdan: Multi-level dependent attention network for visual emotion analysis," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 9479–9488. 2

[13] J. Yang, Q. Huang, T. Ding, D. Lischinski, D. Cohen-Or, and H. Huang, "Emoset: A large-scale visual emotion dataset with rich attributes," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 383–20 394. 1, 2, 4, 6

[14] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proceedings of the International Conference on Machine Learning*, 2023. 1, 2, 3, 7

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[16] D. She, J. Yang, M.-M. Cheng, Y.-K. Lai, P. L. Rosin, and L. Wang, "Wscnet: Weakly supervised coupled networks for visual sentiment classification and detection," *IEEE Transactions on Multimedia*, vol. 22, no. 5, pp. 1358–1371, 2019. 2

[17] S. Zhao, Z. Jia, H. Chen, L. Li, G. Ding, and K. Keutzer, "Pdanet: Polarity-consistent deep attention network for fine-grained visual emotion regression," in *Proceedings of the 27th ACM International Conference on Multimedia*, 2019, pp. 192–201. 2

[18] H. Zhang and M. Xu, "Weakly supervised emotion intensity prediction for recognition of emotions in images," *IEEE Transactions on Multimedia*, vol. 23, pp. 2033–2044, 2020.

[19] W. Zhang, X. He, and W. Lu, "Exploring discriminative representations for image emotion recognition with cnns," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 515–523, 2019. 2

[20] Z. Lian, L. Sun, M. Xu, H. Sun, K. Xu, Z. Wen, S. Chen, B. Liu, and J. Tao, "Explainable multimodal emotion reasoning," *arXiv preprint arXiv:2306.15401*, 2023. 1

[21] J. Wei, M. Bosma, V. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," in *International Conference on Learning Representations*. 1, 3

[22] H. Xie, M.-X. Lee, T.-J. Chen, H.-J. Chen, H.-I. Liu, H.-H. Shuai, and W.-H. Cheng, "Most important person-guided dual-branch cross-patch attention for group affect recognition," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 598–20 608. 1

[23] S. Zhang, L. Dong, X. Li, S. Zhang, X. Sun, S. Wang, J. Li, R. Hu, T. Zhang, F. Wu *et al.*, "Instruction tuning for large language models: A survey," *arXiv preprint arXiv:2308.10792*, 2023. 2

[24] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *Proceedings of the European Conference on Computer Vision*. Springer, 2022, pp. 709–727. 3

[25] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023. 3, 5

[26] M. Dehghani, J. Djolonga, B. Mustafa, P. Padlewski, J. Heek, J. Gilmer, A. P. Steiner, M. Caron, R. Geirhos, I. Alabdulmohsin *et al.*, "Scaling vision transformers to 22 billion parameters," in *International Conference on Machine Learning*. PMLR, 2023, pp. 7480–7512.

[27] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, "Au-assisted graph attention convolutional network for micro-expression recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2871–2880.

[28] ——, "An overview of facial micro-expression analysis: Data, methodology and challenge," *IEEE Transactions on Affective Computing*, 2022.

[29] OpenAI, "Gpt-4 technical report," Tech. Rep., 2023. 1, 2, 4

[30] Z. Xu, Y. Shen, and L. Huang, "Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning," *arXiv preprint arXiv:2212.10773*, 2022. 7, 2

[31] D. Li, J. Li, H. Le, G. Wang, S. Savarese, and S. C. Hoi, "Lavis: A library for language-vision intelligence," *arXiv preprint arXiv:2209.09019*, 2022. 6

[32] E. Parliament, "Eu ai act: first regulation on artificial intelligence," *Accessed June*, vol. 25, p. 2023, 2023. 1

[33] R. Li, S. Sun, M. Elhoseiny, and P. Torr, "Oxfordtvg-hic: Can machine make humorous captions from images?" in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023, pp. 20 293–20 303. 6, 8

[34] K.-C. Peng, T. Chen, A. Sadovnik, and A. C. Gallagher, "A mixed bag of emotions: Model, predict, and transfer emotion distributions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 860–868. 6

[35] Q. You, J. Luo, H. Jin, and J. Yang, "Building a large scale dataset for image emotion recognition: The fine print and the benchmark," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, no. 1, 2016. 2, 6

[36] J. Machajdik and A. Hanbury, "Affective image classification using features inspired by psychology and art theory," in *Proceedings of the 18th ACM international conference on Multimedia*, 2010, pp. 83–92. 6

[37] J. A. Mikels, B. L. Fredrickson, G. R. Larkin, C. M. Lindberg, S. J. Maglio, and P. A. Reuter-Lorenz, "Emotional category data on images from the international affective picture system," *Behavior research methods*, vol. 37, pp. 626–630, 2005. 6

[38] K.-C. Peng, A. Sadovnik, A. Gallagher, and T. Chen, "Where do emotions come from? predicting the emotion stimuli map," in *2016 IEEE international conference on image processing (ICIP)*. IEEE, 2016, pp. 614–618. 6

[39] H. Xie, H. Chung, H.-H. Shuai, and W.-H. Cheng, "Learning to prompt for vision-language emotion recognition," in *2023 11th International Conference on Affective Computing and Intelligent Interaction Workshops and Demos (ACIIW)*, 2023, pp. 1–4. 1